

Chapter 1

Model Fitting Functions and Optimization

In the previous chapter we described the Mx package and outlined the functions therein that may be used to fit a model to data. We now describe these fitting functions in greater detail, and with special emphasis on the genetic analysis of twin data. We also discuss how the fitting functions are used to generate optimal estimates of genetic parameters in the twin design. Again, understanding of these issues is not essential for the use of Mx to fit models to data; we provide this Chapter for those readers wishing to further their knowledge of *how* Mx does what it does, and the statistical theory behind the choice of fit functions for particular types of data.

1.1 Introduction

Testing a theory regarding the causes of human variation requires using the theory to generate a hypothesis, translating the hypothesis into explicit mathematical expressions, and fitting these expressions to observed data. Getting from theory to model is referred to as *model building*; fitting the model to the observed data is referred to as *model fitting*. It is important to realize that theory, hypotheses implied by theory, and the mathematical expressions representing the hypotheses are all prerequisites to actual model fitting. Furthermore, the acceptability of a given model depends on how well it meets four criteria (Eaves *et al.*, 1989a) . A good model should:

- fit the data
- be consistent
- be simple

- have parameters that are statistically significant

An ill-fitting model implies that the theory which generated the model is wrong or requires modification. Consistency means that the model does not violate theoretical constraints. For example, a model that predicts genetic dominance variation in the absence of additive genetic variation is extremely unlikely (see Falconer, 1990, Figure 8.1). Simple models are to be preferred to complex models because they are easier to falsify and they are more informative. Finally, the parameters in the model must be statistically significant. A parameter that does not deviate significantly from zero should be removed from the model and this in turn should have consequences for the theory.

Considering these criteria, we note that the results from model fitting should “feedback” to the theory and the activity of model building. We discussed the relationship between model fitting and model building in some detail in Section ?? of Chapter ??, particularly Figure ??. The interpretation of model failure (or success) is part of the elaboration/decision process that leads to new theories.

1.2 Fitting Models to Data

The minimal requirements for model fitting include a hypothesis upon which to base the model and observed data to test the model. In genetic covariance structure analysis, hypotheses concern the genetic and environmental contributions to phenotypic variance and covariance. For example, one may suppose that individual differences in some continuously varying character, P , are due to additive genetic (G) and unshared environmental influences (E). We translate this hypothesis into the simple model for subject i :

$$P_i = aG_i + eE_i$$

where P_i , G_i , and E_i all represent deviation scores or measurements from the mean values; that is, the means of these variables are zero. The symbols a and e represent the regression coefficients of the phenotype P on the standardized variables G and E . This model implies the following decomposition of the phenotypic variance:

$$\begin{aligned} \text{var}(P) &= a^2\text{var}(G) + e^2\text{var}(E) \\ &= a^2 \times 1 + e^2 \times 1 \\ &= a^2 + e^2, \end{aligned}$$

if $\text{var}(G) = \text{var}(E) = 1$.

To test this model, one collects measurements of the phenotype in a representative sample according to some research design which, under given assumptions, enables one to identify (i.e., uniquely estimate) the regression coefficients a and e . In the twin method, for example, one obtains data from representative samples of

MZ and DZ twins (e.g. Falconer, 1990, Chapter 10). The measurements obtained in these samples are summarized in covariance matrices

\mathbf{S}_{MZ} and \mathbf{S}_{DZ} :

$$\begin{aligned}\mathbf{S}_{\text{MZ}} &= \begin{pmatrix} \text{var}(P_{\text{MZ1}}) & \\ \text{cov}(P_{\text{MZ1}}, P_{\text{MZ2}}) & \text{var}(P_{\text{MZ2}}) \end{pmatrix} \\ \mathbf{S}_{\text{DZ}} &= \begin{pmatrix} \text{var}(P_{\text{DZ1}}) & \\ \text{cov}(P_{\text{DZ1}}, P_{\text{DZ2}}) & \text{var}(P_{\text{DZ2}}) \end{pmatrix}\end{aligned}$$

For each of the elements in these covariance matrices, we can write a mathematical expression in which the observed statistics, the covariance and variances, are related to a number of known and unknown parameters. The parameters are collected in the vector Θ , which can be partitioned into a vector of known (fixed) parameters, Θ_{fi} , and a vector of unknown (free) parameters, Θ_{fr} . Thus, $\Theta = (\Theta'_{fi}, \Theta'_{fr})$. The expected covariances, Σ_{MZ} and Σ_{DZ} , are functions of these parameters:

$$\begin{aligned}\Sigma_{\text{MZ}} &= \begin{pmatrix} a^2\text{var}(G) + e^2\text{var}(E) & \\ a^2\text{var}(G) & a^2\text{var}(G) + e^2\text{var}(E) \end{pmatrix} \\ &= \begin{pmatrix} a^2 + e^2 & \\ a^2 & a^2 + e^2 \end{pmatrix}\end{aligned}$$

and

$$\begin{aligned}\Sigma_{\text{DZ}} &= \begin{pmatrix} a^2\text{var}(G) + e^2\text{var}(E) & \\ .5a^2\text{var}(G) & a^2\text{var}(G) + e^2\text{var}(E) \end{pmatrix} \\ &= \begin{pmatrix} a^2 + e^2 & \\ .5a^2 & a^2 + e^2 \end{pmatrix}\end{aligned}$$

so we have $\Theta' = (\Theta'_{fi}, \Theta'_{fr})$, where

$$\Theta_{fi} = (r_A(\text{MZ}), r_A(\text{DZ}), \text{var}(G), \text{var}(E))',$$

and

$$\Theta_{fr} = (a, e)',$$

in which $r_A(\text{MZ})$ and $r_A(\text{DZ})$ represent correlations among additive genetic values for MZ and DZ twins. The parameters in Θ_{fi} are known on theoretical or statistical grounds. The variances of the latent additive and environmental variables, G and E , are standardized because it is not possible to identify the variances of what are essentially unobserved variables (e.g., Long, 1983). The additive genetic correlation of the MZ twin is unity because the members of a MZ twin pair are genetically identical. The additive genetic correlation of the DZ twin pairs equals .5 under the assumption of random mating. Thus, we arrive at values for the fixed or known parameters, $\Theta_{fi} = (1.0, 0.5, 1.0, 1.0)'$.

Now, a model fitting function is defined as some function of the differences between the expected statistics and the observed statistics, and optimization is concerned with finding values for the unknown parameters, Θ_{fr} , that minimize these differences. In our example, the expected statistics, which depend on the value the vector Θ , are collected in the expected covariance matrices Σ_{MZ} and Σ_{DZ} . We wish to find estimates for a and e such that the discrepancies $\Sigma_{MZ} - \mathbf{S}_{MZ}$ and $\Sigma_{DZ} - \mathbf{S}_{DZ}$ are minimal. We can then consider the *size* of these differences; if they are small we can conclude that the data support the hypothesis; if the differences are large, the hypothesis is rejected. Use of an appropriate function enables us to statistically quantify the terms ‘large’ and ‘small’ in this context.

1.3 Weighted Least Squares Fitting Functions

We will consider various ways of defining fitting functions to quantify the aforementioned discrepancies, which depend on the unknown vector Θ_{fr} . As above, we assume that the observed data are summarized in dispersion matrices (e.g., covariance matrices).

The best known functions are the Unweighted Least Squares (ULS), Generalized Least Squared (GLS), and Maximum Likelihood (ML), all of which are available in Mx (see Chapter ??). These are all special cases of a more general function called Weighted Least Squares (WLS), which we consider first because of its generality and intuitive appeal. Subsequently we will examine the well-known formulations of the ULS, GLS, and ML functions, before returning to WLS.

Let \mathbf{s} and $\boldsymbol{\sigma}$ contain the non-duplicate elements (i.e., the diagonal and sub-diagonal elements) of the observed covariance matrix \mathbf{S} and the model matrix Σ , respectively. The order of \mathbf{S} is k , so that \mathbf{S} and Σ have $\frac{1}{2}k(k+1)$ non-duplicate elements, and the vector \mathbf{s} (and, of course, $\boldsymbol{\sigma}$) is $q = (k \times 1)$ dimensional. For example, the vector $\boldsymbol{\sigma}$ of the matrix Σ_{MZ} is

$$\begin{aligned}\boldsymbol{\sigma} &= \begin{pmatrix} \Sigma_{MZ_{(1,1)}} \\ \Sigma_{MZ_{(2,1)}} \\ \Sigma_{MZ_{(2,2)}} \end{pmatrix} \\ &= \begin{pmatrix} a^2 + e^2 \\ a^2 \\ a^2 + e^2 \end{pmatrix}\end{aligned}$$

where $\Sigma_{MZ_{(i,j)}}$ is the element in the i^{th} row and j^{th} column of the matrix Σ_{MZ} . Let \mathbf{W} denote a $(q \times q)$ positive definite symmetric matrix, where $q = \frac{1}{2}k(k+1)$. The most general formulation is given by the WLS function:

$$F(\theta) = (\mathbf{s} - \boldsymbol{\sigma})' \mathbf{W}^{-1} (\mathbf{s} - \boldsymbol{\sigma}) \quad (1.1)$$

$$= \sum_{g=1}^k \sum_{h=1}^g \sum_{i=1}^k \sum_{j=1}^i w^{gh,ij} (s_{gh} - \sigma_{gh})(s_{ij} - \sigma_{ij}),$$

where

$$\mathbf{s}' = (s_{11}, s_{21}, s_{22}, s_{31}, \dots, s_{kk}),$$

is a vector of the elements in the lower half, including the diagonal, of the covariance matrix \mathbf{S} used to fit the model to the data;

$$\sigma' = (\sigma_{11}, \sigma_{21}, \sigma_{22}, \sigma_{31}, \dots, \sigma_{kk}),$$

is the vector of corresponding elements of $\Sigma(\theta)$ reproduced from the model parameters θ , and $w^{gh,ij}$ is a typical element of the matrix \mathbf{W}^{-1} .

For example, if we let $k = 3$, $\mathbf{d}' = (s_1 - \sigma_1, s_2 - \sigma_2, s_3 - \sigma_3)'$, and we represent a particular element in \mathbf{W}^{-1} by $\mathbf{W}_{i,j}^{-1}$ ($i = 1, 3$ and $j = 1, 3$), we have:

$$\begin{aligned} F(\text{WLS}) &= d_1^2 \mathbf{W}_{1,1}^{-1} + 2d_1 d_2 \mathbf{W}_{2,1}^{-1} + d_2^2 \mathbf{W}_{2,2}^{-1} + 2d_1 d_3 \mathbf{W}_{3,1}^{-1} + \\ &\quad 2d_2 d_3 \mathbf{W}_{3,2}^{-1} + d_3^2 \mathbf{W}_{3,3}^{-1} \end{aligned} \quad (1.2)$$

(remember $\mathbf{W}_{2,1}^{-1} = \mathbf{W}_{1,2}^{-1}$ because \mathbf{W}^{-1} is symmetric). The rationale of the function is easy to see: the discrepancies between the observed and the model statistic are squared and weighted by the inverse of some positive definite symmetric matrix \mathbf{W} . It is the choice of this matrix that determines the fitting function. The ideal selection of \mathbf{W} is one that takes into account the precision and the variances and covariances of the estimates in $(\mathbf{d} = \mathbf{s} - \sigma)$. It is also easy to see that when the fit is perfect this function is zero, because then the elements of \mathbf{d} are all zero.

To obtain consistent parameter estimates, any positive definite matrix \mathbf{W} may be used. Under very general assumptions, if the model holds in the population and if the sample variances and covariances in \mathbf{S} converge in probability to the corresponding elements in the population covariance matrix Σ as the sample size increases, any fit function with a positive definite \mathbf{W} will give a consistent estimator of θ . In practice, the numerical results obtained with one fit function often are so close to those of another that we get the same substantive interpretations of the results.

Analysis of Multiple Groups

In the preceding chapter we noted that genetic covariance structure analyses are, more often than not, multiple group analyses. In the simple example given here we have two covariance matrices, \mathbf{S}_{MZ} and \mathbf{S}_{DZ} . Generally, to obtain a multi-group fitting function, that of each group is weighted by its number of cases and summed:

$$F = (1/N) \sum_{g=1}^G N_g F_g, \quad (1.3)$$

where G is the number of groups, N_g represents the number of cases in group g , F_g the value of the fitting function calculated in group g , and N is the total number of cases, $N = \sum_{g=1}^G N_g$. For example, to obtain a multi-group WLS function we calculate

$$F(\text{WLS}) = (1/N) \sum_{g=1}^G N_g [(\mathbf{s}_g - \boldsymbol{\sigma}_g)' \mathbf{W}_g^{-1} (\mathbf{s}_g - \boldsymbol{\sigma}_g)]. \quad (1.4)$$

The generalized formulation in (1.4) of the fitting function (1.1) has the advantage that it is conceptually easy to understand as a least squares function. However, it is computationally inconvenient because the order of \mathbf{W} can become prohibitively large as the size of \mathbf{W} increases rapidly with k , demanding enormous amounts of computer memory when k is at all large. Consequently, the most common fitting functions mentioned above are not written in this format, but in a more efficient one that we turn to now. To facilitate this discussion, we assume temporarily that the data

1. vary continuously (opposed to, e.g., dichotomous data),
2. are summarized in a covariance matrix (as opposed to, say, a correlation matrix),
3. follow a multivariate normal distribution.

The third assumption is very important because several of the alternative formulations of the fitting function are possible only under this assumption. Each method and its characteristics will be discussed individually. Then, these assumptions will be relaxed and we will return to the WLS function.

1.3.1 Unweighted Least Squares Fitting Function (ULS)

The ULS function is defined as

$$\begin{aligned} F(\text{ULS}) &= (\mathbf{s} - \boldsymbol{\sigma})' \mathbf{I}^{-1} (\mathbf{s} - \boldsymbol{\sigma}) \\ &= (\mathbf{s} - \boldsymbol{\sigma})' (\mathbf{s} - \boldsymbol{\sigma}), \end{aligned} \quad (1.5)$$

or in its more familiar form:

$$F(\text{ULS}) = \frac{1}{2} \text{tr}[(\mathbf{S} - \boldsymbol{\Sigma})^2]. \quad (1.6)$$

Note that the weight matrix used in ULS is an identity matrix ($\mathbf{W} = \mathbf{I}$). Thus, this function minimizes one-half the sum of squares of each of the elements in the residual matrix $\mathbf{S} - \boldsymbol{\Sigma}$. Finding the values for Θ_{fr} that minimize this function yields ULS parameter estimates, $\hat{\Theta}(\text{ULS})$. It can be verified easily that $F(\text{ULS})$ has as its theoretical minimum zero. The ULS function has advantages and disadvantages. Its advantages are:

1. It has good convergence properties; that is to say it is not difficult to minimize.
2. It is simple to program and computationally undemanding.
3. It yields consistent parameter estimates, so as the sample size increases (or, more generally, the information that the estimator uses increases), the estimates $\hat{\Theta}(\text{ULS})$ converge to Θ_{fr} stochastically with probability 1.0. This characteristic is not conditional on distributional assumptions.
4. It will work under a wide variety of conditions, including non-positive definite states of the input or model covariance matrices.

Its disadvantages are:

1. No statistical inference. It does not generally yield a statistic to rigorously test the overall goodness of fit of the specified model. Therefore, once estimates have been obtained, there is no formal criterion to decide whether or not to reject the specified hypothesis.
2. The parameter estimates are *scale dependent* (see Browne, 1982; Krane and McDonald, 1978). Let \mathbf{D} be a diagonal matrix of positive scale factors. Scale dependence implies that the results obtained by analyzing \mathbf{S} and those obtained by analyzing \mathbf{DSD}' are not properly related. Given one solution and knowledge of the matrix \mathbf{D} , one cannot derive the other solution. This is a serious disadvantage: if two researchers investigate identical phenotypes with different scales, their results and conclusions regarding the quantitative genetic decomposition of the phenotypic variance may differ!
3. Standard errors of the estimates are not generally available.
4. The ULS estimator is not as efficient as others. That is to say, *ceteris paribus*, the precision of the ULS estimates is not as great as that obtained by the other methods.

The problems associated with ULS stem from the fact that all the discrepancies $\mathbf{S} - \Sigma$ contribute equally to the fitting function (the choice of $\mathbf{W} = \mathbf{I}$ leads to this equal weighting). In other words, no consideration is given to the fact that there might be differences in the precision with which an element in \mathbf{S} is estimated (Joreskog, 1988). It is desirable to take into account such differences by weighting the discrepancies in a manner that reflects the precision and interdependence of the estimates in \mathbf{S} . The next fitting function, GLS, carries out just such a weighting.

1.3.2 Generalized Least Squares (GLS)

The GLS function has the following form:

$$F(\text{GLS}) = (\mathbf{s} - \sigma)' \mathbf{W}_{(\text{GLS})}^{-1} (\mathbf{s} - \sigma), \quad (1.7)$$

and a more convenient form:

$$F(\text{GLS}) = \frac{1}{2} \text{tr}[(\mathbf{I} - \mathbf{S}^{-1}\Sigma)^2]. \quad (1.8)$$

The matrix $\mathbf{W}(\text{GLS})$ in (1.7) contains the covariances and variances of the terms in \mathbf{s} . Due to the assumption of multivariate normality many of these terms are known (in the same sense that the skewness and kurtosis of a normally distributed variable are known) and the more convenient form in (1.8) results. We now see that the matrix \mathbf{S}^{-1} functions as a weight matrix for the residual matrix $\mathbf{S} - \Sigma$. The theoretical minimum is zero, as can be readily verified. Under the distributional assumption noted above, this estimator has a number of advantages over the ULS function. The advantages of GLS are

1. Scale invariance (in the sense defined above: $F(\text{GLS})[\mathbf{S}; \Sigma] = F(\text{GLS})[\mathbf{DSD}; \mathbf{D}\Sigma\mathbf{D}]$).
2. The asymptotic covariance matrix of the estimates $\hat{\Theta}(\text{GLS})$ is known so that standard errors are available.
3. An overall goodness of fit test statistic is available (chi-squared) so that one may have a formal criterion to determine the tenability of the hypothesis (statistical inference).
4. The GLS estimator is consistent and has greater efficiency than ULS.

A clear disadvantage of GLS is that the violation of the distributional assumptions lead to incorrect overall test statistic and standard errors. Also, \mathbf{S} must be positive definite, although this can hardly be called a disadvantage.

1.3.3 Maximum Likelihood (ML)

The ML function (or likelihood ratio function) is perhaps the most frequently used fitting function. It is defined as

$$F(\text{ML}) = (\mathbf{s} - \sigma)' \mathbf{W}_{(\text{ML})}^{-1} (\mathbf{s} - \sigma), \quad (1.9)$$

and in its more convenient form as

$$F(\text{ML}) = \frac{1}{2} \text{tr}[(\mathbf{S} - \Sigma)\Sigma^{-1}]^2. \quad (1.10)$$

Finally, in a more familiar form:

$$F(\text{ML}) = \log |\Sigma| + \text{tr}(\mathbf{S}\Sigma^{-1}) - \log |\mathbf{S}| - p, \quad (1.11)$$

where p is the order of input covariance matrix \mathbf{S} . The latter formulation is less obvious than the two equivalents. Bollen (1989, Appendix 4A & 4B) gives an

excellent derivation of this formulation of the ML function. The matrix $\mathbf{W}_{(ML)}^{-1}$ now contains the inverse of the variances and covariances among the elements in the matrix Σ .

The advantages of this fitting function are the same as the GLS function and, like GLS, the major disadvantage is the distributional assumption of multivariate normality. Compared to GLS, ML has the computational disadvantage that Σ has to be inverted at each iteration, but aside from this these fitting functions are very similar (Browne, 1974).

1.3.4 Fitting Functions Including Means

The fit functions we have discussed so far are for covariance structures *only*; they are augmented if we fit models to observed means as well as covariances (see Section ?? and Chapter ??). When means are involved, the fit functions are augmented by a term of the form:

$$F(\text{mean}) = \frac{1}{2}(\mathbf{z} - \boldsymbol{\mu})' \mathbf{W}^{-1}(\mathbf{z} - \boldsymbol{\mu}) \quad (1.12)$$

where \mathbf{z} and $\boldsymbol{\mu}$ are respectively the vectors of observed and expected means, and \mathbf{W}^{-1} is the observed covariance matrix for ULS, GLS, DWLS and WLS, but is the expected covariance matrix for ML. Note that just like the General WLS function for covariances, equation 1.12 evaluates to zero when the observed and expected means are equal. Thus when $\mathbf{z} = \boldsymbol{\mu}$ we are left with a function of the covariance structure alone.

1.3.5 Weighted Least Squares (WLS) Revisited

So far we have been concerned with normally distributed continuously varying variables and we have assumed that the data are summarized in a covariance matrix. We now return to the general and useful fitting function called the weighted least squares function, WLS. The WLS function can be used in many situations; for example, when the data are summarized in a correlation matrix, when the data follow an arbitrary distribution, or when the data are discrete (dichotomous or polychotomous). In (1.1) we presented the WLS function

$$F(\text{WLS}) = (\mathbf{s} - \boldsymbol{\sigma})' \mathbf{W}^{-1}(\mathbf{s} - \boldsymbol{\sigma}) \quad (1.13)$$

This function is quite general because it can accommodate, by suitable choice of a \mathbf{W} matrix, many types of data summary (correlation, covariance), and data following almost any distribution. The issue is to choose the appropriate \mathbf{W} to suit the data summary or their distributional characteristics or both.

The Weight Matrix, \mathbf{W}

In Section 1.3 we noted that if the model of the data is correct in the population and the elements of \mathbf{S} converge to those in Σ as the sample size increases, any fit

function will give a consistent estimator of Θ , given a positive definite \mathbf{W} . We also have noted that the usual way of choosing \mathbf{W} in weighted least squares is to let any element of \mathbf{W} , $w_{gh,ij}$, be a consistent estimate of the asymptotic covariance between s_{gh} and s_{ij} . If this is the case, we say that \mathbf{W}^{-1} is a *correct weight matrix*. However, further assumptions must be made if one needs an asymptotically correct chi-squared measure of goodness-of-fit and asymptotically correct standard errors of parameter estimates.

“Classical” theory for covariance structures (see e.g., Browne, 1974, or Jöreskog, 1981), assumes that the variances and covariances of the elements of \mathbf{S} are of the form

$$\text{ACov}(s_{gh}, s_{ij}) = (1/N)(\sigma_{gi}\sigma_{hj} + \sigma_{gj}\sigma_{hi}) . \quad (1.14)$$

This holds, in particular, if the observed variables have a multivariate normal distribution, or if \mathbf{S} has a Wishart distribution. The GLS and ML methods and their chi-square values and standard errors are based on (1.14). The GLS method corresponds to using a matrix \mathbf{W}^{-1} in (1.1) whose general element is

$$w^{gh,ij} = N(2 - \delta_{gh})(2 - \delta_{ij})(s^{gi}s^{hj} + s^{gj}s^{hi}) , \quad (1.15)$$

where δ_{gh} and δ_{ij} are Kronecker deltas¹. The fit function (1.11) for ML is not of the form (1.1) but may be shown to be equivalent to using a \mathbf{W}^{-1} of the form (1.15), with s replaced by an estimate of σ which is updated in each iteration.

In recent fundamental work by Browne (1982, 1984), this classical theory for covariance structures has been generalized to any multivariate distribution (including non-normal distributions) for continuous variables satisfying very mild assumptions. This approach uses a \mathbf{W} matrix with typical element

$$w_{gh,ij} = m_{ghij} - s_{gh}s_{ij} , \quad (1.16)$$

where

$$m_{ghij} = (1/N) \sum_{a=1}^N (z_{ag} - \bar{z}_g)(z_{ah} - \bar{z}_h)(z_{ai} - \bar{z}_i)(z_{aj} - \bar{z}_j) \quad (1.17)$$

are the fourth-order central moments. Using such a \mathbf{W} in (1.1) gives what Browne calls “asymptotically distribution free (ADF)” estimators, for which correct asymptotic chi-squares and standard errors may be obtained. Browne has shown that this \mathbf{W} matrix also may be used to compute correct asymptotic chi-squares and standard errors for estimates which have been obtained by the classical ML and GLS methods. WLS uses \mathbf{W} as defined by (1.16), whereas GLS uses the \mathbf{W} formulation in (1.15), which shows that WLS and GLS are different forms of weighted least squares: *WLS is asymptotically distribution free, while GLS is based on normal theory.*

¹The value of the Kronecker delta δ_{ij} is zero for $i \neq j$ and unity for $i = j$.

The advantages of the WLS function include all those of ML and GLS, in addition to the advantage of postulating virtually no distributional assumptions. A considerable disadvantage is that very large sample sizes are required to arrive at a reliable estimate of \mathbf{W} . This means that the asymptotic characteristics of the distribution of the fitting function and covariance matrix of the estimates become trustworthy only when the sample size is large. A second drawback is the computational aspects of the function. We noted previously that the \mathbf{W} matrix quickly becomes very large as the number of input variables (i.e., the order of the input covariance matrix) increases. If we have k variables to analyze, we have $q = \frac{1}{2}k(k+1)$ non-duplicate elements in \mathbf{S} and $u = \frac{1}{2}q(q+1)$ non-duplicate elements in \mathbf{W} . Thus, for example, analysis of $k = 20$ variables would yield a covariance matrix, \mathbf{S} , with $q = 210$ elements, and a weight matrix \mathbf{W} with $u = 22155$ unique entries! A third drawback is that when there are missing observations in the data, different moments involved in (1.16) may be based on different numbers of cases unless listwise deletion is used. When pairwise deletion is used, it is not clear how to deal with this problem.

1.3.6 Additional Fitting Functions: Modified ML, DWLS, and Functions for Raw Data

Three further functions warrant a brief mention. First, there is a modified ML fitting function which yields correct overall χ^2 goodness-of-fit statistics and standard errors when the data are non-normally distributed and the assumption is made that the marginal distribution of the phenotypic variables are characterized by an identical kurtosis (the multivariate distribution is then called an “elliptical distribution”). When this hypothesis cannot be rejected, one can use the ML fitting function and correct the results (i.e., χ^2 goodness of fit index and standard errors) for the departure from normality (see Bollen, 1989; Browne, 1984). This method is very attractive, but the assumption of an elliptical distribution is rather restrictive.

The second fitting function is Diagonally Weighted Least Squares (DWLS; Jöreskog and Sörbom, 1989), which involves using only the diagonal elements of the \mathbf{W} matrix when analyzing covariance matrices calculated from non-normal continuously varying variables. This function gives correct standard errors and χ^2 only under very limited circumstances. It is used when the WLS option is required, but the full \mathbf{W} matrix does not fit in the computer memory; for example, in analyses of large correlation matrices. The DWLS function is considered to be a compromise between ULS and (ADF) WLS. It does not seem to have much to recommend it, although it is superior to ULS.

Finally, we mention the possibility of fitting models to raw data. So far we have assumed that the phenotypic data are summarized in one or more dispersion matrices. This method of data summarization is very convenient but requires that each case (or “pedigree”) has the same structure. When the data comprise cases of variable composition this convenient method of data summarization cannot be

used for practical and statistical reasons and the model has to be fit directly to the raw data (Lange, *et al.*, 1976). A pedigree may vary in composition from a single individual to a complex pedigree comprising many different social and biological relationships. When the structure of the cases is highly variable (the pedigrees are said to be unbalanced), it is possible to use Maximum Likelihood by calculating the log-likelihood for each pedigree separately:

$$L_i = -\frac{1}{2} \log |\Sigma_i| - \frac{1}{2} (\mathbf{x} - \mathcal{E}[\mathbf{x}])' \Sigma_i^{-1} (\mathbf{x} - \mathcal{E}[\mathbf{x}]) + \text{constant}, \quad (1.18)$$

where Σ_i is the expected covariance matrix for the i^{th} pedigree, \mathbf{x} is the vector of observed data within that pedigree, and $\mathcal{E}[\mathbf{x}]$ is a vector of expected means. The constant in this expression is $j \log(2\pi)$, where j is the number of variables being analyzed. This log-likelihood is maximized over all of the pedigrees $L = \sum_{i=1}^k L_i$, where k is the number of distinct pedigrees. This method is available in the package Mx (Neale, 1991). It is useful to recognize that when there are no missing data (the data are “balanced pedigrees”) the sum of the k log-likelihoods in equation 1.18 can be expressed as (e.g., Mardia *et al.*, 1979, p. 97):

$$L_A = -\frac{k}{2} \log |\Sigma_i| - \frac{k}{2} \text{tr}(\Sigma^{-1} \mathbf{S}) - \frac{k}{2} (\bar{\mathbf{x}} - \mu)' \Sigma_i^{-1} (\bar{\mathbf{x}} - \mu) + \text{constant}. \quad (1.19)$$

In large populations, $\bar{\mathbf{x}} = \mu$, so the term $(\bar{\mathbf{x}} - \mu)' \Sigma_i^{-1} (\bar{\mathbf{x}} - \mu)$ is zero, thus reducing to

$$L_A = -\frac{k}{2} \log |\Sigma_i| - \frac{k}{2} \text{tr}(\Sigma^{-1} \mathbf{S}) + \text{constant}. \quad (1.20)$$

For an explanation of the very close relationship between between the log-likelihood ratio function and the likelihood function, the reader is referred to Bollen (1989) or Lawley and Maxwell (1971). Neale *et al.*, (1989b) also have discussed this relationship in the context of genetic model-fitting.

Figure 1.1 serves as a partial summary of the preceding discussion of fitting functions and may be used as an aid in determining a suitable function when analyzing dispersion matrices.

1.3.7 Goodness of Fit and the Principle of Parsimony

Each of the fit-functions described up to now measures the overall goodness of fit of a model. However, this really addresses only the first of the four aspects of model-fitting described in the introduction to this Chapter, namely whether or not the model fits the data. The mathematical consistency of the model is addressed by the use of path analysis and structural equations, but we have not considered how to judge whether or not a model is *simple*.

Quite often, simply adding more and more parameters to a model will allow it to explain (i.e., fit) the data perfectly. However, as we saw in Section ??, a model that

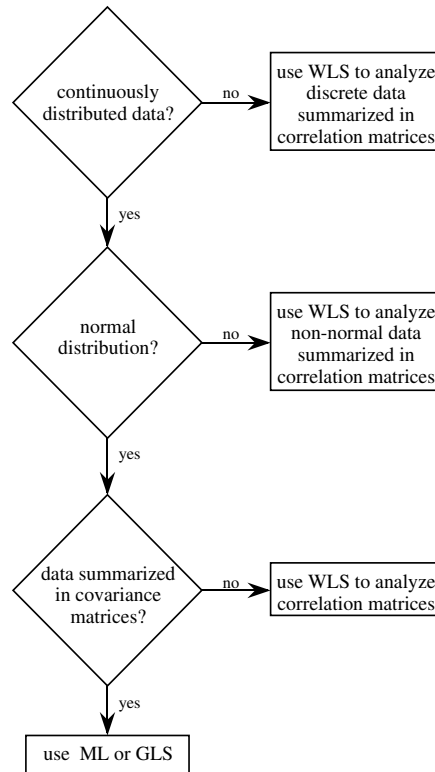


Figure 1.1: Flow chart to aid selection of a fitting function for the analysis of dispersion matrices.

has to fit perfectly does little more than transform the data. While it may prove useful for interpretation of data, it has the disadvantage of not being *falsifiable*. Thus a model that fits well is not necessarily useful; what counts is the ability to predict a wide range of phenomena with a smaller set of parameters. This is the principle of parsimony, recognized by the fourteenth-century philosopher, William of Occam, to whom we owe the term *Occam's razor*. For practical purposes, we can get an idea of the parsimony of a model by combining its goodness-of-fit χ^2 with its degrees of freedom. There are many such indices of fit (see, e.g., Browne and Cudeck, 1988; Marsh *et al.* 1988; Bollen 1989, Mulaik *et al* 1989; Kaplan, 1990) and the one we have selected here should not be considered to be necessarily the best, but for a variety of genetic models it seems to agree with expert opinion most of the time. . We compute Akaike's Information Criterion for a model as

$$\chi^2 - 2df$$

and the model with the lowest (i.e., largest negative) value of this index is said to fit best by AIC. Fitting best by this criterion is clearly *not* the same as fitting best by the χ^2 statistic; therefore we caution authors against the unqualified use of expressions such as ‘model II fits best.’

1.4 Optimization

1.4.1 Introduction

We stated earlier that optimization is concerned with finding values for the unknown parameters, Θ_{fr} , that minimize a fitting function. Now we shall try to give some idea of how optimization works. We limit the discussion to a method known as the quasi-Newton algorithm. This and other methods are explained fully in Dennis and Schnabel (1983) and in Gill, Murray and Wright (1981).

As above, we have a fitting function F and a vector containing unknown parameters Θ . We estimate these parameters by finding values for them that minimize our chosen F . In general, minimization requires an iterative process because the parameter estimates that minimize the function cannot be solved in closed form. One well-known iterative method of minimizing a function is *Newton-Raphson*. We will consider this approach briefly in order to introduce the related quasi-Newton method and to introduce some terminology.

Consider the following simple optimization problem. The correlation for the personality trait “dominance” is .53 for MZ twins and .25 for DZ twins (Loehlin and Nichols, 1976). Suppose that we hypothesize that additive genetic effects are responsible for these correlations. This hypothesis leads to the pair of (slightly inconsistent) simultaneous equations: $.53 = a^2$ and $.25 = .5a^2$, which we will fit using the ULS function. As mentioned previously, in multiple group designs the loss function is the weighted sum of loss functions calculated in each group, $(1/N) \sum_{i=1}^G N_g F_g(\text{ULS})$, where N_g and F_g are respectively the sample size and function value for group g , and $N = \sum_{i=1}^G N_g$. So we have

$$F(\text{ULS}) = \frac{N_{MZ}}{N_{MZ} + N_{DZ}} (a^2 - .53)^2 + \frac{N_{DZ}}{N_{MZ} + N_{DZ}} (.5a^2 - .25)^2,$$

where N_{MZ} equals the number of MZ twin pairs ($N_{MZ} = 490$) and N_{DZ} equals the number of DZ twin pairs ($N_{DZ} = 317$). Then,

$$F(\text{ULS}) = .6072(a^2 - .53)^2 + .3928(.5a^2 - .25)^2.$$

To apply the Newton-Raphson minimization method, we require two sets of information. The first is the so-called first-order derivative of the function with respect to the unknown parameter a . This also is referred to as the *gradient*, and is represented by the m dimensional gradient vector \mathbf{g} (m equals the number of unknown parameters). For example,

$$\begin{aligned}\frac{\partial F}{\partial a} &= \frac{1}{N_{\text{MZ}} + N_{\text{DZ}}} a [a^2(4N_{\text{MZ}} + N_{\text{DZ}}) - (2.12N_{\text{MZ}} + .50N_{\text{DZ}})] \\ &= 2.8216a^3 - 1.4837a\end{aligned}$$

so $\mathbf{g}' = [\frac{\partial F}{\partial a}]'$. Because in the present example we have only one parameter to estimate ($m = 1$), the vector \mathbf{g} is one dimensional. The second piece of information is the matrix of second order partial derivatives. This ($m \times m$) symmetric matrix is called the *Hessian*, \mathbf{H} , and is in the present example

$$\begin{aligned}\frac{\partial^2 F}{\partial a \partial a'} &= \frac{1}{N_{\text{MZ}} + N_{\text{DZ}}} 3a^2(4N_{\text{MZ}} + N_{\text{DZ}}) - \\ &\quad (2.12N_{\text{MZ}} + .50N_{\text{DZ}})N_{\text{MZ}} + N_{\text{DZ}} \\ &= 8.4648a^2 - 1.4837.\end{aligned}$$

The Newton-Raphson algorithm is an iterative scheme that works as follows:

1. Choose a starting value for the vector Θ , the unknown parameters. Call this vector $\Theta^{(k)}$.
2. Calculate the function value, gradient vector, and Hessian matrix: $F^{(k)}$, $\mathbf{g}^{(k)}$, and $\mathbf{H}^{(k)}$.
3. Calculate the direction vector $\mathbf{d}^{(k)} = \mathbf{H}^{-1(k)} \mathbf{g}^{(k)}$.
4. Calculate $\Theta^{(k+1)} = \Theta^{(k)} - \mathbf{d}^{(k)}$; i.e., the estimates for iteration $k + 1$.
5. Goto (2) replacing $\Theta^{(k)}$ by $\Theta^{(k+1)}$.

Given the starting values $\Theta^{(k)}$ this algorithm determines the direction in which to proceed by calculating the quantity $\mathbf{d}^{(k)}$. This continues until the difference between $F^{(k+1)}$ and $F^{(k)}$ is smaller than some predetermined constant, ϵ (e.g., $\epsilon = 0.0001$). Applying this algorithm to the example ULS fitting function starting at the point $a = 1.0$ yields the following results:

Iteration	a	F	\mathbf{g}	\mathbf{H}	$\mathbf{H}^{-1}\mathbf{g}$
1	1.0	.158681	1.3379360	6.981136	.1916502
2	.808350	1.156227E-02	.2910483	4.047485	7.190842E-02
3	.736441	2.686400E-04	3.433225E-02	3.107186	1.104931E-02
4	.725392	7.617289E-05	7.572186E-04	2.970459	2.549163E-04
5	.725137	7.607631E-05	3.625686E-07	2.967330	1.221868E-07
6	.725137	7.607632E-05	0	2.967328	0

Thus, the estimate $\hat{a} = .725$ is obtained from minimizing this fitting function. We note that \mathbf{g} is zero at the minimum and that \mathbf{H} is positive (2.96). These are necessary and sufficient conditions for the point $a = .725$ to be a true (local) minimum (Yamane, 1968). In the case that we have more than one unknown parameter these same conditions apply: the vector \mathbf{g} should be a zero vector and the Hessian \mathbf{H} should be positive definite.

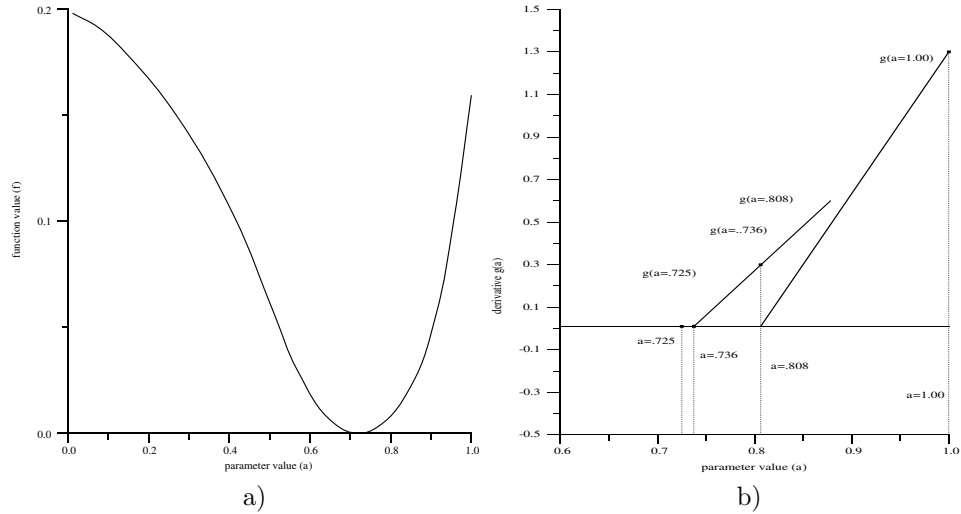


Figure 1.2: Graphic depiction of a) parameter space, and b) iterative progression of Newton-Raphson algorithm.

Figure 1.2a depicts the function values given various values of the parameter a . The function can be seen to be approximately quadratic in the vicinity of the minimum ($a = .725$). Figure 1.2b demonstrates graphically the progression during the first two iterations of the Newton Raphson algorithm.

This algorithm generally works well given good starting values, but requires exact (analytic) expressions for both the first (\mathbf{g}) and the second order derivatives (\mathbf{H}). The latter can be time consuming and complicated to calculate for the fitting functions used in covariance structure analysis. The quasi-Newton method uses the same approach, but the exact Hessian is replaced by some approximation, \mathbf{Y} . Both \mathbf{g} and \mathbf{H} can be estimated by finite differences, but the Hessian can be approximated by a number of additional methods. The choice of the method, \mathbf{Y} , determines the optimization routine. The use of a matrix other than the exact Hessian complicates the Newton-Raphson algorithm in two respects:

1. A choice of \mathbf{Y} has to be made. \mathbf{Y} can be chosen to be fixed (unchanging throughout iteration) or can be updated during iteration (a new \mathbf{Y} for each new set of estimates). The latter requires an additional algorithm for \mathbf{Y} .

2. An algorithm has to be determined at step (4) given above, because the calculation of $\Theta^{(k+1)} = \Theta^{(k)} - \mathbf{H}^{-1(k)}\mathbf{g}^{(k)}$ is replaced by $\Theta^{(k+1)} = \Theta^{(k)} - \alpha\mathbf{Y}^{-1(k)}\Theta^{(k)}$ where α is an unknown quantity called the *step length parameter* because it determines how much the current value of $\Theta^{(k+1)}$ will be changed using the information $\mathbf{Y}^{-1(k)}\mathbf{g}^{(k)}$.

1.4.2 Choice of Hessian Approximation

It is mainly the choice of \mathbf{Y} that determines the quasi-Newton optimization algorithm. The following gives an overview of possibilities:

Choice of \mathbf{Y}	Algorithm
$\mathbf{Y} = \mathbf{H}$. Exact Hessian	Newton-Raphson
$\mathbf{Y} \approx \mathbf{H}$. Finite Difference Approximation	Discrete Newton
Information Matrix	Gauss-Newton or Fisher's Scoring Method
$\mathbf{Y} = \mathbf{I}$. Identity matrix	Steepest Descent

Mx uses the Davidon-Fletcher-Powell (DFP) method as the default means of determining \mathbf{Y} . In this method, parameter estimates, gradients, and \mathbf{Y} at iteration k , are used to obtain an approximation of \mathbf{Y} for iteration $k + 1$ (see e.g., Lawley and Maxwell, 1971, for the relevant formulas). As iteration proceeds, the DFP update of \mathbf{Y} converges on the matrix \mathbf{H} (Gill, *et al.* 1981). Other options in Mx involve calculating the information matrix at each iteration. The information matrix is the Hessian calculated under the assumption that the true model is being fit [$\mathcal{E}(\mathbf{S}) = \cdot$]. Although this may seem like a very strong assumption, in practice it has been found to work well (Dolan and Molenaar, 1991). When one chooses the information matrix for \mathbf{Y} , the algorithm is called the Fisher-scoring method. Jöreskog (1988) points out that the Fisher scoring method and the Gauss-Newton algorithm are identical. Finally, one can choose a matrix \mathbf{Y} that is fixed throughout iteration. For example, the steepest descent algorithm involves replacing \mathbf{Y} with the identity matrix, \mathbf{I} .

1.5 Summary

In this chapter we have described and compared some of the commonly used fitting functions for structural equation modeling, including the group of weighted least squares and maximum likelihood functions for fitting to moment data matrices,

and the maximum likelihood pedigree approach of Lange *et al.* (1976) for fitting to raw data. The ML moment matrix procedure is the most frequently used function for genetic analysis of twin data (it is also the default fit function in Mx), due in large part to its desirable properties for estimators, standard errors, and relation to the chi-squared goodness-of-fit and likelihood ratio tests, and the fact that it has been reasonably well characterized in the literature. However, each of the fitting functions described here has unique advantages, which permits structural analysis of twin data in nearly any form or composition.

We also have presented a brief overview of optimization theory and practice, with emphasis upon the Newton-Raphson and quasi-Newton algorithms. Our description of these procedures has been kept simple, even superficial, in order to illustrate how parameters are estimated under a chosen model, yet not divert attention from our primary focus of genetic analysis of twin and family data. One of the chief advantages of using a commercial software package such as Mx is that one need not be burdened by the intricacies of optimization, which are legion, but may, instead, concentrate on the characteristics of the data themselves. The remainder of this book is directed toward this task.